Project Submission Guidelines

- The project is due on December 10 at 11:59pm.
- This is a group project. It can be done in groups of 2 or 3.
- There should be **only one** report submission (PDF) and **only one** notebook (Python notebook) submission **per group** through Gradescope.
  - Please write down the names of all group members on the report.
  - Each group should submit a notebook containing all the code. Note that this notebook should not simply contain the codes, but all the relevant text explaining what is done for what purpose, comments useful for readability of the code, and outputs including visualizations accompanying each code cell. Make sure the notebook you provide is able to access all the data necessary.
  - Further details regarding the reports and the notebook files can be found in the Project Grading section.

Project Topic

The main topic of the project is crime in Tucson. We want to inspect factors that may be correlated and/or contributing to occurrences of Tucson crimes and the types thereof. You will need to carefully consider the ethical implications of your analysis, including your selection of data and methods as well as the possibility of unintended consequences when predictive policing tools are applied in the real world.

Project Goal

Primary goal is to analyze and model crime in the Tucson area using data available at City of Tucson Data Hub and other relevant data sources helpful to predict crime, understand crime's effect on society, or understand the effects of predictive policing on society.

Expected Tasks

This is an open-ended project in the sense that each group is responsible to create their own questions/hypotheses of interest and the prediction/classification task(s) of the study. You are expected to do the following:

- Forming questions/hypotheses of interest
- Data gathering
- Data preprocessing (merging different data sources, data cleaning, etc.)
- Exploratory data analysis and visualizations
- Revising questions/hypotheses of interest (if necessary) and explicitly stating the prediction/classification task(s)
- Implementation of **at least two models** for the defined task(s)
- Appropriate evaluation of the models
- Communicating the results (report and notebook)

Data Sources

Useful crimes/arrests data sources:

- Click here for Tucson Police Reported Crimes.
- Click here for Police Arrests 2021. (Includes more detailed location information than the Crimes data above. Can be obtained for years prior to 2021 as well.)

Useful data sources to consider together with the crimes/arrests data:

- Click here for streetlight locations.
- Click here for neighborhood income.

  Click here for parcel - regional data. (Although the website marks it as deprecated, it includes more detailed location information than the neighborhood income data above and also includes full cash value of the listed parcels.)

You are expected to search for other relevant data at City of Tucson Data Hub: Click here for the Data Hub. You may add relevant data from sources other than the Data Hub as well.

Project Grading

- **Grading Points for the notebooks (70pts):**
  - Data (including retrieval, cleaning, exploratory data analysis and visualization):
    - **If any one of the following holds [15pts]:** Choice of datasets is not appropriate for the aimed questions/hypotheses. Data cleaning procedure is not sufficient or is inappropriate or is missing though necessary. Analysis/visualization is overly simplistic or incomplete.
    - **If the following holds [23pts]:** All the relevant items above are handled appropriately, though there is still room for improvement in at least one of the aspects.
    - **If the following holds [30pts]:** Choice of datasets, cleaning procedure, or analysis/visualization is complete, advanced, and informative.
  - Model (including data preprocessing, model building, evaluation framework, result reporting and visualization):
    - **If any one of the following holds [15pts]:** Data preprocessing not handled properly, choice of model(s) is not appropriate, the evaluations are not proper, the evaluation framework is not appropriately implemented, conclusions are missing or incorrect, choice of plots of results inappropriate/missing.
    - **If the following holds [23pts]:** All the relevant items above are handled appropriately, though there is still room for improvement in at least one of them.
    - **If the following holds [30pts]:** All the relevant items above are handled completely, leaving little room for improvement in any one of them.
  - Notebook readability:
    - **If any one of the following holds [5pts]:** The texts explaining relevant notebook cells are inappropriate or missing. Code is messy and poorly organized; unused or irrelevant code distracts when reading code. Variables and functions names are not helpful to understand code. Comments are inappropriate or missing.
    - **If the following holds [8pts]:** All the relevant items above are reasonable, though there is still room for improvement in at least one of them.
    - **If the following holds [10pts]:** The texts explaining relevant notebook cells are perfectly explanatory. Code is very well organized. No irrelevant or distracting code. Variable and function names have clear relationship to their purpose in the code. Code is easy to read and understand. Very-well commented.

- **Grading Points for the reports (30pts):**

  The report should consist of the following sections:

  - Introduction: The definition of the problem should be provided and the questions/hypothesis to be considered should be introduced.
  - Related Work: Relevant literature search should be provided. There should be a brief summary of existing similar systems and the main novelties your proposed system/technique/analysis/tool brings over the existing systems.
  - Methods: The overall framework of the project should be described (describe the pipeline, the models employed, the evaluation framework). The employed methods should be discussed in detail; the appropriateness of the models, the parameters etc.
  - Results: A discussion of the main findings supported with nice evaluation plots and visualizations etc.

- Conclusion: Provide concluding remarks. Discuss any future relevant work that might be interesting to pursue.
- References: Provide references to the related work as well as to the datasets employed etc. Make sure within the text of the report you cite the reference at the appropriate places.

For grading purposes:

- **If any one of the following holds [15pts]:** Any one of the sections summarized above has a weak content.
- **If the following holds [23pts]:** If all the sections are proper, but there is still room for improvement in at least one of them.
- **If the following holds [30pts]:** All the relevant sections above are handled completely, leaving little room for improvement in any one of them.